1. 29 approved SMARTS string of toxicophore

|  |  |  |  |
| --- | --- | --- | --- |
|  | Toxicophore name | SMARTS string | Excluded SMARTS string |
| 1 | specific arom nitro | O=N(~O)a | O=N(O)c[$(aS(=O)=O),$(aaS(=O)=O),$(aaaS(=O)=O),$(aC((F)F)F),$(aaC((F)F)F),$(aaaC((F)F)F)] |
| 2 | specific arom amine | a[NH2] | [NH2]a[$(a[$(C((F)F)F),$(S(=O)=O),$(C(=O)O)]),$(aa[$(C((F)F)F),$(S(=O)=O),$(C(=O)O)]),$(aaa[$(C((F)F)F),$(S(=O)=O),$(C(=O)O)])] |
| 3 | aromatic nitroso | a[N;X2]=O |  |
| 4 | alkyl nitrite | CO[N;X2]=O |  |
| 5 | nitrosamine | N[N;X2]=O |  |
| 6 | epoxide | O1[c,C]-[c,C]1 |  |
| 7 | aziridine | C1NC1 |  |
| 8 | azide | N=[N+]=[N-] |  |
| 9 | diazo | C=[N+]=[N-] |  |
| 10 | triazene | N=N-N |  |
| 11 | aromatic azo | c[N;X2]!@;=[N;X2]c | [$([N;X2]([$(acS((=O)=O)),$(aacS((=O)=O)),$(aaacS((=O)=O)),$(aaaacS((=O)=O))])=[N;X2][$(acS((=O)=O)),$(aacS((=O)=O)),$(aaacS((=O)=O)),$(aaaacS((=O)=O))])] |
| 12 | unsubstituted heteroatom-bonded heteroatom | [OH,NH2][N,O] | O=N(O)[O-] |
| 13 | aromatic hydroxylamine | [OH]Na |  |
| 14 | aliphatic halide | [Cl,Br,I]C |  |
| 15 | carboxylic acid halide | [Cl,Br,I]C=O |  |
| 16 | nitrogen or sulfur mustard | [N,S]!@[C;X4]!@[CH2][Cl,Br,I] |  |
| 17 | bay-region in polycyclic aromatic hydrocarbons | [cH]1[cH]ccc2c1c3c(cc2)cc[cH][cH]3 |  |
| 18 | K-region in polycyclic aromatic hydrocarbons | [cH]1cccc2c1[cH][cH]c3c2ccc[cH]3 |  |
| 19 | polycyclic aromatic system | [$(a13~a~a~a~a2~a1~a(~a~a~a~3)~a~a~a~2),$(a1~a~a~a2~a~1~a~a3~a(~a~2)~a~a~a~3),$(a1~a~a~a2~a~1~a~a~a3~a~2~a~a~a~3),$(a1~a~a~a~a2~a~1~a3~a(~a~2)~a~a~a~a~3),$(a1~a~a~a~a2~a~1~a~a3~a(~a~2)~a~a~a~3),$(a1~a~a~a~a2~a~1~a~a3~a(~a~2)~a~a~a~a~3),$(a1~a~a~a~a2~a~1~a~a~a3~a~2~a~a~a~3),$(a1~a~a~a~a2~a~1~a~a~a3~a~2~a~a~a~a~3),$(a13~a~a~a~a2~a1~a(~a~a~a~3)~a~a~2)] |  |
| 20 | sulfonate-bonded carbon (alkyl alkane sulfonate or dialkyl sulfate) | [$([C,c]OS((=O)=O)O!@[c,C]),$([c,C]S((=O)=O)O!@[c,C])] |  |
| 21 | aliphatic N-nitro | O=N(~O)N |  |
| 22 | R,â-unsaturated aldehyde (including R-carbonyl aldehyde) | [$(O=[CH]C=C),$(O=[CH]C=O)] | [$(O=[CH]C([N,O,S])=C),$(O=[CH]C=C[N,O,S]),$(O=[CH]C=Ca)] |
| 23 | diazonium | [N;v4]#N |  |
| 24 | â-propiolactone | O=C1CCO1 |  |
| 25 | R,â-unsaturated alkoxy group | [CH]=[CH]O |  |
| 26 | 1-aryl-2-monoalkyl hydrazine | [NH;!R][NH;!R]a |  |
| 27 | aromatic methylamine | [CH3][NH]a | [CH3][NH]a[$(a[$(C((F)F)F),$(S=O),$(C(=O)O)]),$(aa[$(C((F)F)F),$(S=O),$(C(=O)O)]),$(aaa[$(C((F)F)F),$(S=O),$(C(=O)O)])] |
| 28 | ester derivative of aromatic hydroxylamine | aN([$([OH]),$(O\*=O)])[$([#1]),$(C(=O)[CH3]),$([CH3]),$([OH]),$(O\*=O)] |  |
| 29 | polycyclic planar system | [$([X2,X3]13~[X2,X3]~[X2,X3]~[X2,X3]~[X2,X3]2~[X2,X3]1~[X2,X3](~[X2,X3]~[X2,X3]~[X2,X3]~3)~[X2,X3]~[X2,X3]~[X2,X3]~2),$([X2,X3]1~[X2,X3]~[X2,X3]~[X2,X3]2~[X2,X3]~1~[X2,X3]~[X2,X3]3~[X2,X3](~[X2,X3]~2)~[X2,X3]~[X2,X3]~[X2,X3]~3),$([X2,X3]1~[X2,X3]~[X2,X3]~[X2,X3]2~[X2,X3]~1~[X2,X3]~[X2,X3]~[X2,X3]3~[X2,X3]~2~[X2,X3]~[X2,X3]~[X2,X3]~3),$([X2,X3]1~[X2,X3]~[X2,X3]~[X2,X3]~[X2,X3]2~[X2,X3]~1~[X2,X3]3~[X2,X3](~[X2,X3]~2)~[X2,X3]~[X2,X3]~[X2,X3]~[X2,X3]~3),$([X2,X3]1~[X2,X3]~[X2,X3]~[X2,X3]~[X2,X3]2~[X2,X3]~1~[X2,X3]~[X2,X3]3~[X2,X3](~[X2,X3]~2)~[X2,X3]~[X2,X3]~[X2,X3]~3),$([X2,X3]1~[X2,X3]~[X2,X3]~[X2,X3]~[X2,X3]2~[X2,X3]~1~[X2,X3]~[X2,X3]3~[X2,X3](~[X2,X3]~2)~[X2,X3]~[X2,X3]~[X2,X3]~[X2,X3]~3),$([X2,X3]1~[X2,X3]~[X2,X3]~[X2,X3]~[X2,X3]2~[X2,X3]~1~[X2,X3]~[X2,X3]~[X2,X3]3~[X2,X3]~2~[X2,X3]~[X2,X3]~[X2,X3]~3),$([X2,X3]1~[X2,X3]~[X2,X3]~[X2,X3]~[X2,X3]2~[X2,X3]~1~[X2,X3]~[X2,X3]~[X2,X3]3~[X2,X3]~2~[X2,X3]~[X2,X3]~[X2,X3]~[X2,X3]~3),$([X2,X3]13~[X2,X3]~[X2,X3]~[X2,X3]~[X2,X3]2~[X2,X3]1~[X2,X3](~[X2,X3]~[X2,X3]~[X2,X3]~3)~[X2,X3]~[X2,X3]~2)] |  |

1. The dataset file (in .sdf format) is downloaded from PubChem Malaria Bioassay dataset using AID number (AID: 449704, NOVARTIS: Inhibition of Plasmodium falciparum W2 (drug-resistant) proliferation in erythrocyte-based infection assay - BioAssay Summary) provided in the assignment description. Then choose the “active” under the “Tested Compounds” on the right hand bar, continuing with downloading compound structures in .sdf format.

Here, I used the OpenBabelGUI to implement my filtering. (The GUI is downloaded from the OpenBabel main page and is operated under Windows 7 system.)

On the left hand side, it’s showing the INPUT FORMAT. Choose the input format to be sdf – MDL MOL format. And choose the dataset file (3955862273781340154.sdf) to be the input file.

In the middle part, it’s where we put our filtering requirements: in the “Chemical Object Option(multichar)” , for the blank “Filter: convert only when tests are true”, type in our SMARTS (with s= means search for substructures that match this SMRATS and s!= means excluded the ones, finally put the SMARTS sting inside the two single quote makes ‘SMARTS\_string’).

Ex1: s='O=N(~O)a' s!='O=N(O)c[$(aS(=O)=O),$(aaS(=O)=O),$(aaaS(=O)=O),$(aC((F)F)F),$(aaC((F)F)F),$(aaaC((F)F)F)]'

Ex2: s='a[N;X2]=O'

On the right hand side, it’s about output format. Choose the “smi – SMILES format” option and the output will show up in the column.

|  |  |  |
| --- | --- | --- |
|  | Toxicophore name | Frequency |
| 1 | specific arom nitro | 309 |
| 2 | specific arom amine | 123 |
| 3 | aromatic nitroso | 0 |
| 4 | alkyl nitrite | 0 |
| 5 | nitrosamine | 1 |
| 6 | epoxide | 11 |
| 7 | aziridine | 2 |
| 8 | azide | 0 |
| 9 | diazo | 0 |
| 10 | triazene | 1 |
| 11 | aromatic azo | 5 |
| 12 | unsubstituted heteroatom-bonded heteroatom | 17 |
| 13 | aromatic hydroxylamine | 0 |
| 14 | aliphatic halide | 83 |
| 15 | carboxylic acid halide | 0 |
| 16 | nitrogen or sulfur mustard | 1 |
| 17 | bay-region in polycyclic aromatic hydrocarbons | 3 |
| 18 | K-region in polycyclic aromatic hydrocarbons | 1 |
| 19 | polycyclic aromatic system | 310 |
| 20 | sulfonate-bonded carbon (alkyl alkane sulfonate or dialkyl sulfate) | 4 |
| 21 | aliphatic N-nitro | 1 |
| 22 | α,β-unsaturated aldehyde (including R-carbonyl aldehyde) | 0 |
| 23 | diazonium | 0 |
| 24 | β-propiolactone | 0 |
| 25 | α,β -unsaturated alkoxy group | 1 |
| 26 | 1-aryl-2-monoalkyl hydrazine | 31 |
| 27 | aromatic methylamine | 3 |
| 28 | ester derivative of aromatic hydroxylamine | 0 |
| 29 | polycyclic planar system | 358 |
|  |  |  |

1. The results of filtering compounds using SMARTS and OpenBabel showed the four toxicophores with the top four highest frequency (as highlighted in yellow): specific arom nitro(309), specific arom amine(123), polycyclic aromatic system(310), and polycyclic planar system(358).

For the compounds mentioned in the paper, there are three precursors to produce the aromatic hydroxylamide intermediate: aromatic nitro group with enzymatic reduction, aromatic nitroso with nonenzymatic reduction, and aromatic amide with enzymatic oxidation. With subsequent activation, the intermediate becomes electrophilic and covalently binds to DNA. As a result, this could leads to mutation in the DNA molecules. The aromatic nitro and amide substructure which we regarded as generic toxicophores could raise their accuracy of predicting mutagenicity by identifying their detoxifying substructures and observing their steric hindrance. It is also showed that polycyclic aromatic and planar systems are can act as intercalation agents that interact with DNA.

In summary, using the Malaria Bioassay datasets (AID: 449704), we read back the cheminfo paper published in 2005 and proved 4 toxicophores to be as valid as using the information technology nowadays. These four toxicophores are molecules that have interaction have DNA so we can use them to predict mutagenicity among molecules dataset.